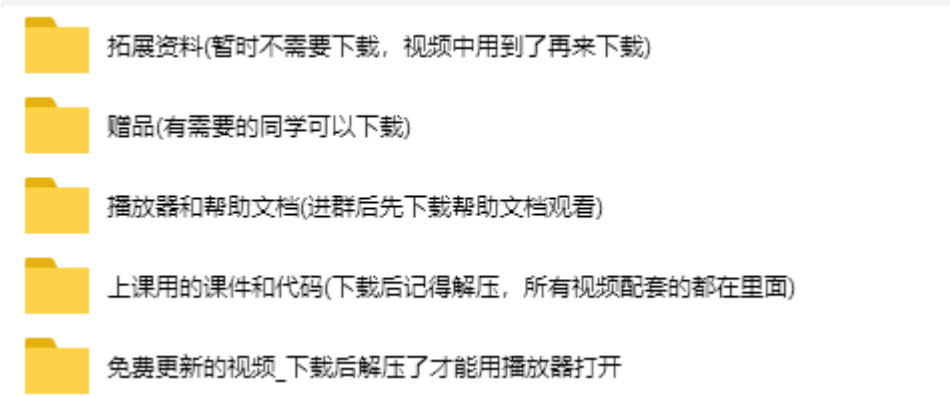
第四讲:拟合算法

与插值问题不同，在拟合问题中不需要曲线一定经过给定的点。拟合问题的目标是寻求一个函数（曲线），使得该曲线在某种准则下与所有的数据点最为接近，即曲线拟合的最好（最小化损失函数）。

温馨提示

（1）视频中提到的附件可在**售后群的群文件**中下载。

包括**讲义、代码、我视频中推荐的资料**等。

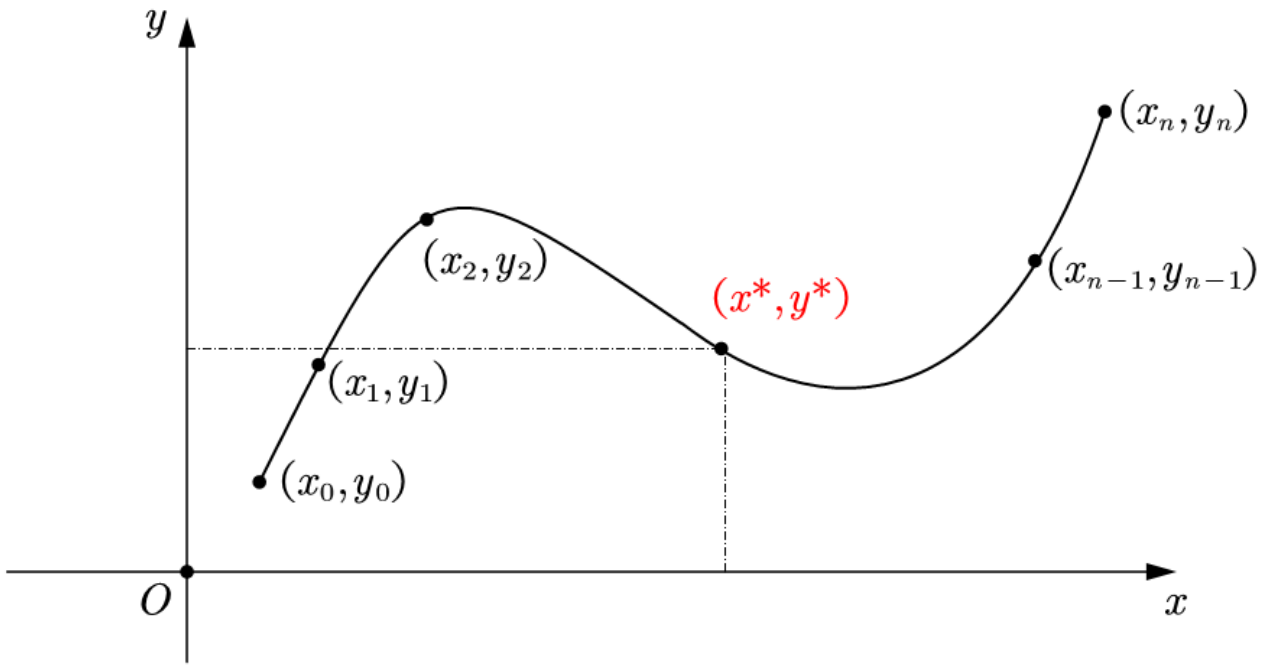


（2）关注我的**微信公众号《数学建模学习交流》**，后台发送**“软件”**两个字，可获得常见的建模软件下载方法；发送**“数据”**两个字，可获得建模数据的获取方法；发送**“画图”**两个字，可获得数学建模中常见的画图方法。另外，也可以看看公众号的历史文章，里面发布的都是对大家有帮助的技巧。

（3）**购买更多优质精选的数学建模资料**，可关注我的微信公众号《数学建模学习交流》，在后台发送**“买”**这个字即可进入店铺进行购买。

（4）视频价格不贵，但价值很高。单人购买观看只需要**58元**，和另外两名队友一起购买人均仅需**46元**，视频本身也是下载到本地观看的，所以请大家**不要侵犯知识产权**，对视频或者资料进行二次销售。

插值和拟合的区别



插值算法中，得到的多项式f(x)要经过所有样本点。但是如果样本点太多，那么这个多项式次数过高，会造成龙格现象。

尽管我们可以选择分段的方法避免这种现象，但是更多时候我们更倾向于得到一个确定的曲线，尽管这条曲线不能经过每一个样本点，但只要保证误差足够小即可，这就是拟合的思想。**(拟合的结果是得到一个确定的曲线)**

一个小例子

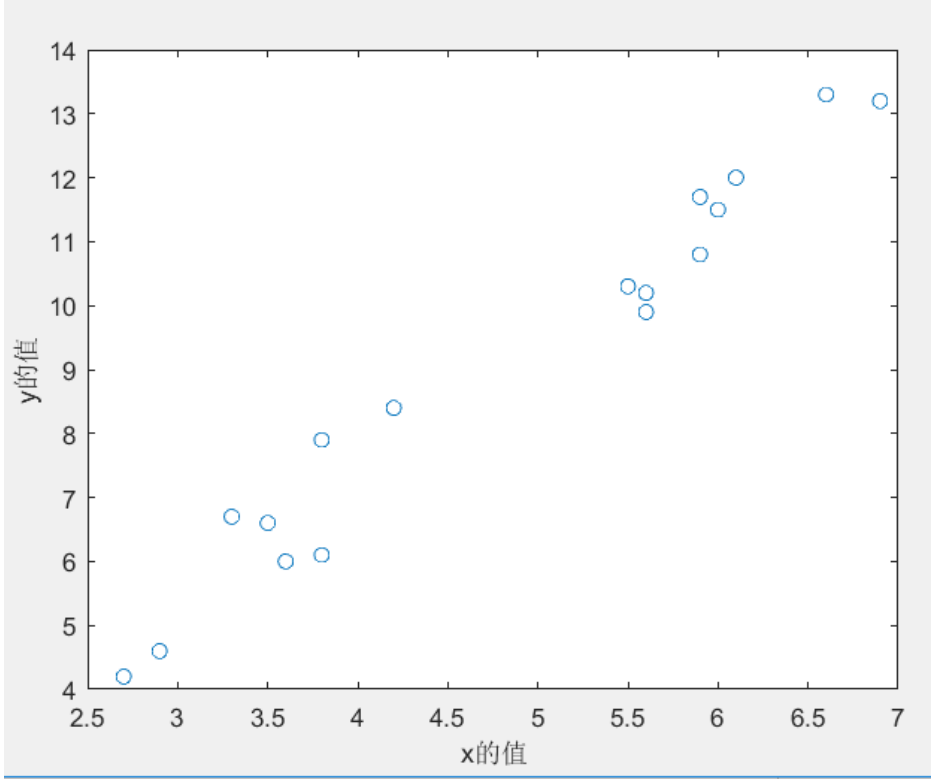
代码文件:code1.m

找出y和x之间的拟合曲线

x	y
4.2	8.4
5.9	11.7
2.7	4.2
3.8	6.1
3.8	7.9
5.6	10.2
6.9	13.2
3.5	6.6
3.6	6
2.9	4.6
4.2	8.4
6.1	12
5.5	10.3
6.6	13.3
2.9	4.6
3.3	6.7
5.9	10.8
6	11.5
5.6	9.9

数据文件：data1.xlsx

```
plot(x,y,'o')
% 给x和y轴加上标签
xlabel('x的值')
ylabel('y的值')
```



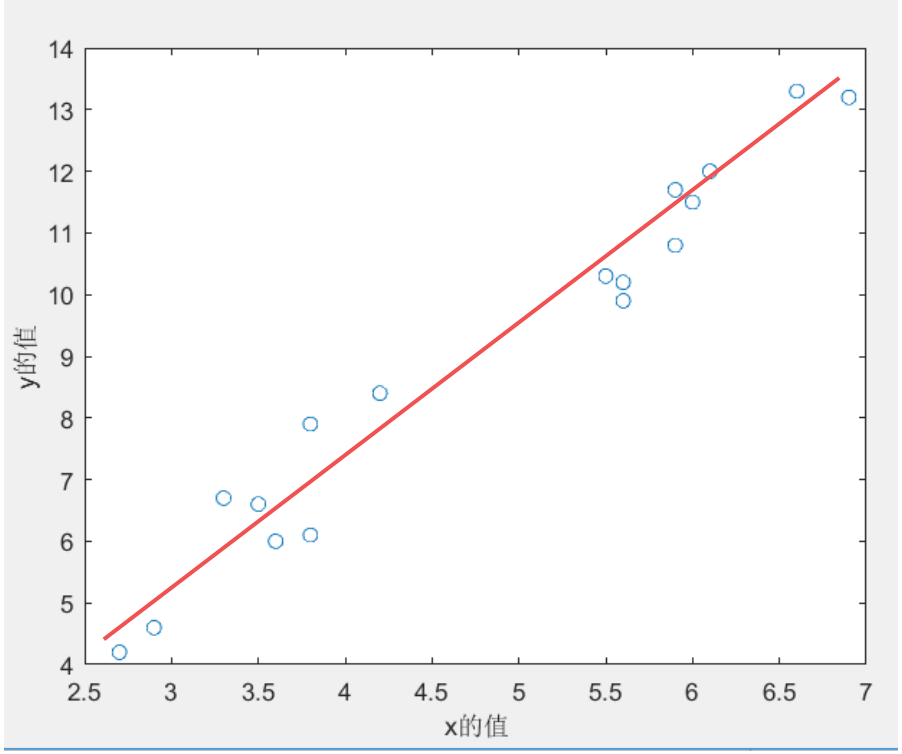
注意：将数据导入到matlab时，我们是分别创建了两个变量：x和y，每个变量仅保存一列数据。

确定拟合曲线

设这些样本点为(x_i, y_i), $i = 1, 2, \cdots, n$

我们设置的拟合曲线为$y = kx + b$

问题：k和b取何值时，样本点和拟合曲线最接近。

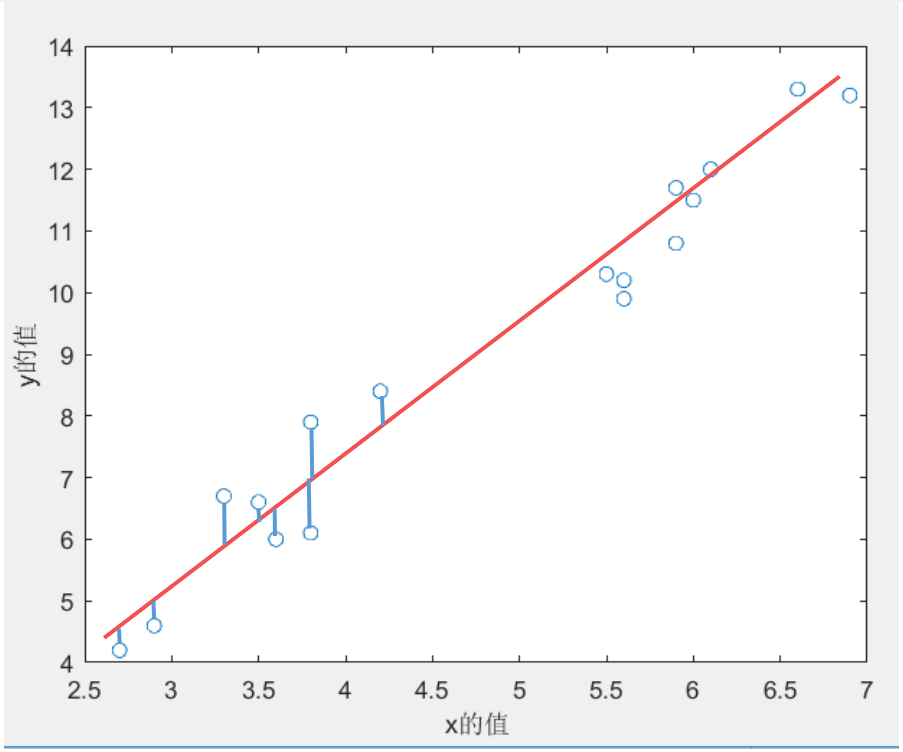


最小二乘法的几何解释

设这些样本点为(x_i, y_i), $i = 1, 2, \cdots, n$

我们设置的拟合曲线为$y = kx + b$

问题：k和b取何值时，样本点和拟合曲线最接近。



第一种定义：$\hat{y}_i = kx_i + b$

$$\hat{k}, \hat{b} = \arg\min_{k,b}\left(\sum_{i=1}^{n}|y_i - \hat{y}_i|\right)$$

第二种定义：$\hat{y}_i = kx_i + b$

$$\hat{k}, \hat{b} = \arg\min_{k,b}\left(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right)$$

第一种定义有绝对值，不容易求导，因此计算比较复杂。
所以我们往往使用第二种定义，这也正是最小二乘的思想。

为什么不用四次方？

不使用三次方的原因：会出现正负相抵的情况
不使用四次方的原因：若存在异常值，其对曲线的干扰较大

（1）避免极端数据对拟合曲线的影响。
（2）最小二乘法得到的结果和MLE极大似然估计一致。
不用奇数次方的原因：误差会正负相抵。

求解最小二乘法

设这些样本点为(x_i,y_i)，$i=1,2,\cdots,n$，我们设置的拟合曲线为$y=kx+b$

令拟合值$\hat{y}_i = kx_i+b$

$$\text{那么}\hat{k},\hat{b}=\underset{k,b}{arg\min}\left(\sum_{i=1}^{n}(y_i-\hat{y}_i)^2\right)=\underset{k,b}{arg\min}\left(\sum_{i=1}^{n}(y_i-kx_i-b)^2\right)$$

$$\text{令}L=\sum_{i=1}^{n}(y_i-kx_i-b)^2\text{，现要找}k,b\text{使得}L\text{最小。}$$

只需要了解即可：　　（L在机器学习中被称为损失函数，在回归中也常被称为残差平方和）

$$\begin{cases}\dfrac{\partial L}{\partial k}=-2\sum_{i=1}^{n}x_i(y_i-kx_i-b)=0\\ \dfrac{\partial L}{\partial b}=-2\sum_{i=1}^{n}(y_i-kx_i-b)=0\end{cases}\Rightarrow\begin{cases}\sum_{i=1}^{n}x_iy_i=k\sum_{i=1}^{n}x_i^2+b\sum_{i=1}^{n}x_i\\ \sum_{i=1}^{n}y_i=k\sum_{i=1}^{n}x_i+bn\end{cases}\Rightarrow\begin{cases}n\sum_{i=1}^{n}x_iy_i=kn\sum_{i=1}^{n}x_i^2+bn\sum_{i=1}^{n}x_i\\ \sum_{i=1}^{n}y_i\sum_{i=1}^{n}x_i=k\sum_{i=1}^{n}x_i\sum_{i=1}^{n}x_i+bn\sum_{i=1}^{n}x_i\end{cases}$$

$$n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}y_i\sum_{i=1}^{n}x_i=kn\sum_{i=1}^{n}x_i^2-k\sum_{i=1}^{n}x_i\sum_{i=1}^{n}x_i\ \Rightarrow\ \hat{k}=\frac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}y_i\sum_{i=1}^{n}x_i}{n\sum_{i=1}^{n}x_i^2-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}x_i}$$

同理我们可得到：$\hat{b}=\dfrac{\sum_{i=1}^{n}x_i^2\sum_{i=1}^{n}y_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}x_iy_i}{n\sum_{i=1}^{n}x_i^2-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}x_i}$

证明不需要了解，知道k和b的结论即可（doge）

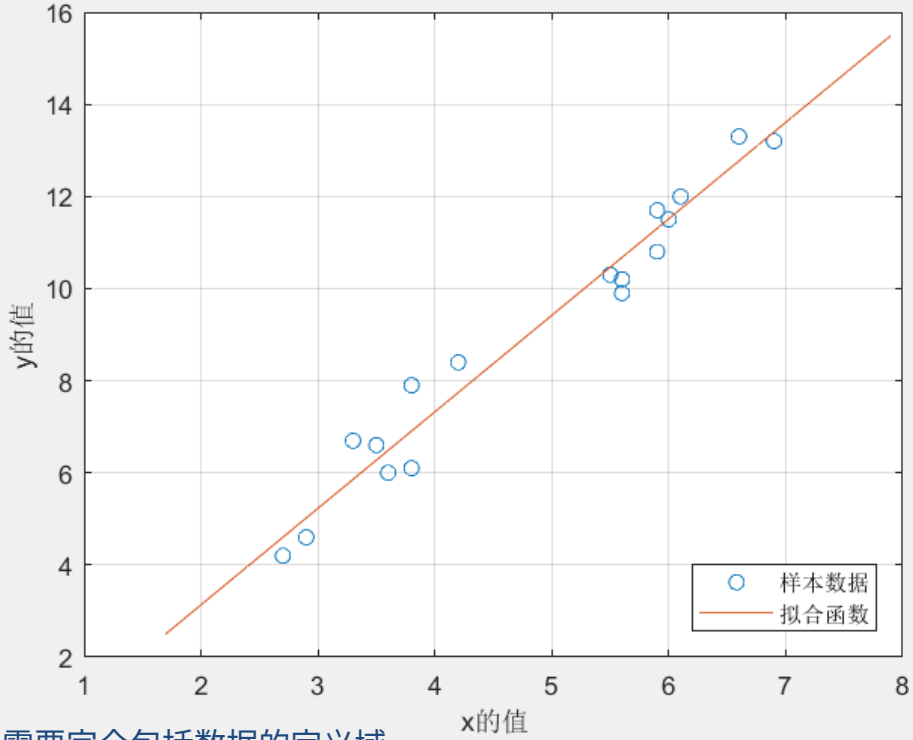
Matlab求解最小二乘

代码文件:code1.m

只需要知道这个

$$\hat{k}=\frac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}y_i\sum_{i=1}^{n}x_i}{n\sum_{i=1}^{n}x_i^2-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}x_i}\quad,\quad \hat{b}=\frac{\sum_{i=1}^{n}x_i^2\sum_{i=1}^{n}y_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}x_iy_i}{n\sum_{i=1}^{n}x_i^2-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}x_i}$$

```
clear;clc
load data1
plot(x,y,'o')
% 给x和y轴加上标签
xlabel('x的值')
ylabel('y的值')
n = size(x,1);
k = (n*sum(x.*y)-sum(x)*sum(y))
/(n*sum(x.*x)-sum(x)*sum(x))
b = (sum(x.*x)*sum(y)-sum(x)*
sum(x.*y))/(n*sum(x.*x)-sum(x)
*sum(x))
hold on % 继续在之前的图形上来画图形
grid on % 显示网格线
f=@(x) k*x+b;
fplot(f,[2.5,7]);
legend('样本数据','拟合函数','location','SouthEast')
```



这是图像的定义域，需要完全包括数据的定义域

另外一种使用plot函数作图，需要自己生成一组x数据，不如fplot函数方便

如何评价拟合的好坏

拟合优度（可决系数）R^2

总体平方和SST：$Total\ sum\ of\ squares$：$SST = \sum_{i=1}^{n}(y_i - \overline{y})^2$

误差平方和SSE：$The\ sum\ of\ squares\ due\ to\ error$：$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

回归平方和SSR：$Sum\ of\ squares\ of\ the\ regression$：$SSR = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$

可以证明：$SST = SSE + SSR$（要用到我们求导得到的两个等式）

拟合优度：$0 \le R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST} \le 1$

R^2越接近1，说明误差平方和越接近0，误差越小说明拟合的越好。

（注意：R^2只能用于拟合函数是线性函数时，拟合结果的评价）

线性函数和其他函数（例如复杂的指数函数）比较拟合的好坏，直接看SSE即可

（未来你可能有机会看到R^2是个负数）

证明SST = SSE + SSR

总体平方和SST：*Total sum of squares*：$SST = \sum_{i=1}^{n}(y_i-\overline{y})^2$

误差平方和SSE：*The sum of squares due to error*：$SSE = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2$

回归平方和SSR：*Sum of squares of the regression*：$SSR = \sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2$

$$\sum_{i=1}^{n}(y_i-\overline{y})^2 = \sum_{i=1}^{n}(y_i-\hat{y}_i+\hat{y}_i-\overline{y})^2 = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2 + 2\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\overline{y})$$

一阶导数条件

$$\begin{cases} \dfrac{\partial L}{\partial k} = -2\sum_{i=1}^{n}x_i(y_i-kx_i-b) = 0 \\ \dfrac{\partial L}{\partial b} = -2\sum_{i=1}^{n}(y_i-kx_i-b) = 0 \end{cases}$$

$$\Rightarrow \begin{cases} \sum_{i=1}^{n}x_i(y_i-\hat{y}_i) = 0 \\ \sum_{i=1}^{n}(y_i-\hat{y}_i) = 0 \end{cases}$$

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\overline{y})$$

$$= \sum_{i=1}^{n}\hat{y}_i(y_i-\hat{y}_i) - \overline{y}\sum_{i=1}^{n}(y_i-\hat{y}_i)$$

$$= \sum_{i=1}^{n}(kx_i+b)(y_i-\hat{y}_i) = 0$$

"线性函数"的介绍

R^2只能用于拟合函数是“线性函数”时，拟合结果的评价

思考：$y = a + bx^2$是线性函数吗？

是的，因为我们这里说的线性函数是指**对参数为线性（线性于参数）**。

由于本书主要讨论像方程（2.2.2）那样的线性模型，所以我们必须知道线性一词的真正含义，因为对它可作两种解释。

□ 对变量为线性

对线性的第一种并且也许是更“自然”的一种解释是，Y 的条件期望值是 X_i 的线性函数，比如说，方程（2.2.2）。① 从几何意义上说，这时回归曲线是一条直线。按照这种解释，诸如 $E(Y \mid X_i) = \beta_1 + \beta_2 X_i^2$ 的回归函数，由于变量 X 以幂或指数 2 出现，就不是线性的。

□ 对参数为线性

对线性的第二种解释是，Y 的条件期望 $E(Y \mid X_i)$ 是参数 β 的一个线性函数；它可以是或不是变量 X 的线性函数。② 对于这种解释，$E(Y \mid X_i) = \beta_1 + \beta_2 X_i^2$ 就是一个线性（于参数）回归模型。为了看出这一点，让我们假设 X 取值为 3。因此，$E(Y \mid X = 3) = \beta_1 + 9\beta_2$，显然它是 β_1 和 β_2 的线性函数。图 2—3 中所示的所有模型因此也都是线性回

参考资料：古扎拉蒂《计量经济学基础》第五版

如何判断线性于参数的函数?

在函数中，参数仅以一次方出现，且不能乘以或除以其他任何的参数，并不能出现参数的复合函数形式。

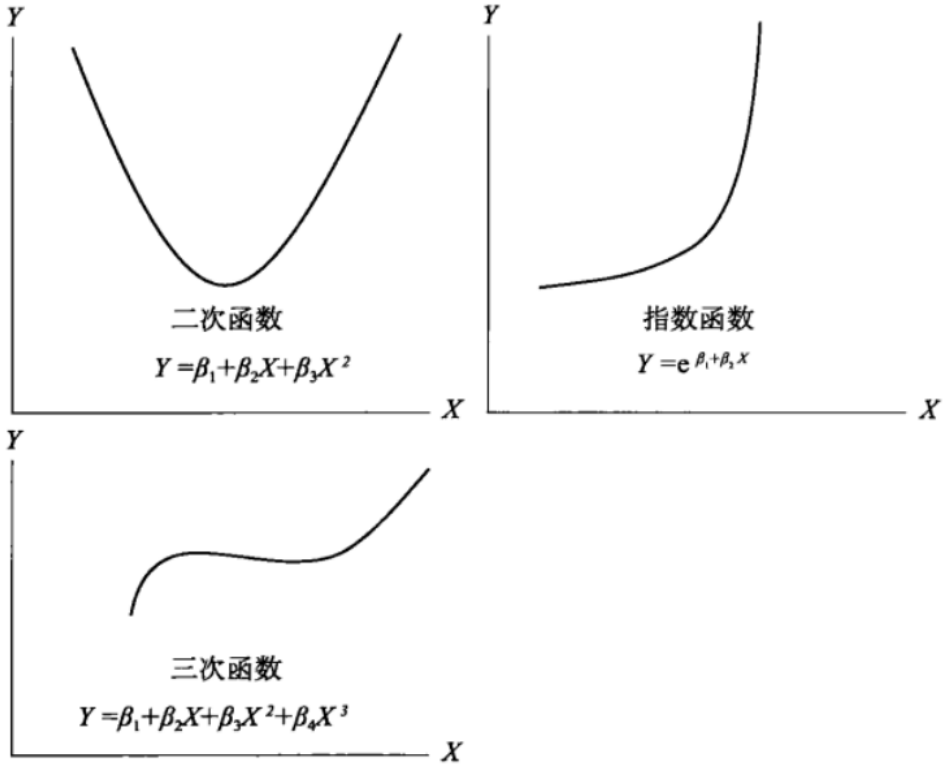


图 2—3　线性于参数的函数

$y=a/(x-b)^2$、$y=a\sin(b+cx)$都不是线性函数，不能用R^2.

计算拟合优度的代码

代码文件:code1.m

$$SST = \sum_{i=1}^{n}(y_i - \overline{y})^2$$

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

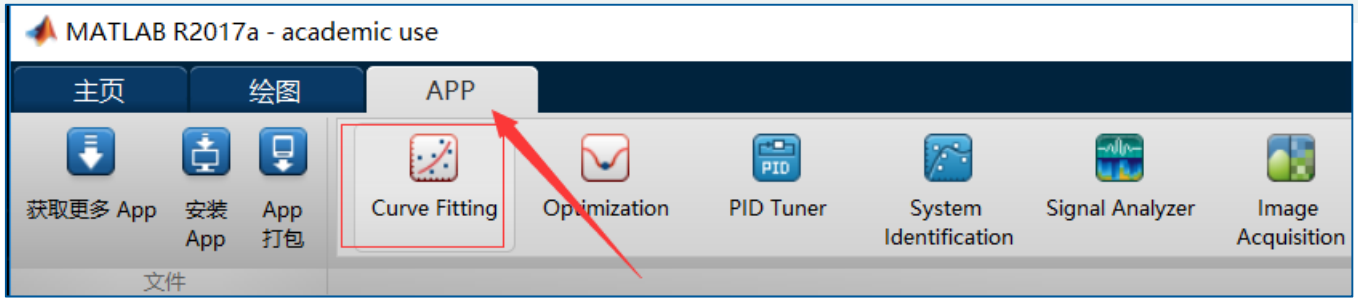
$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$$

```
y_hat = k*x+b;    % y的拟合值
SSR = sum((y_hat-mean(y)).^2)    % 回归平方和
SSE = sum((y_hat-y).^2)          % 误差平方和
SST = sum((y-mean(y)).^2)        % 总体平方和
SST-SSE-SSR
R_2 = SSR / SST
```

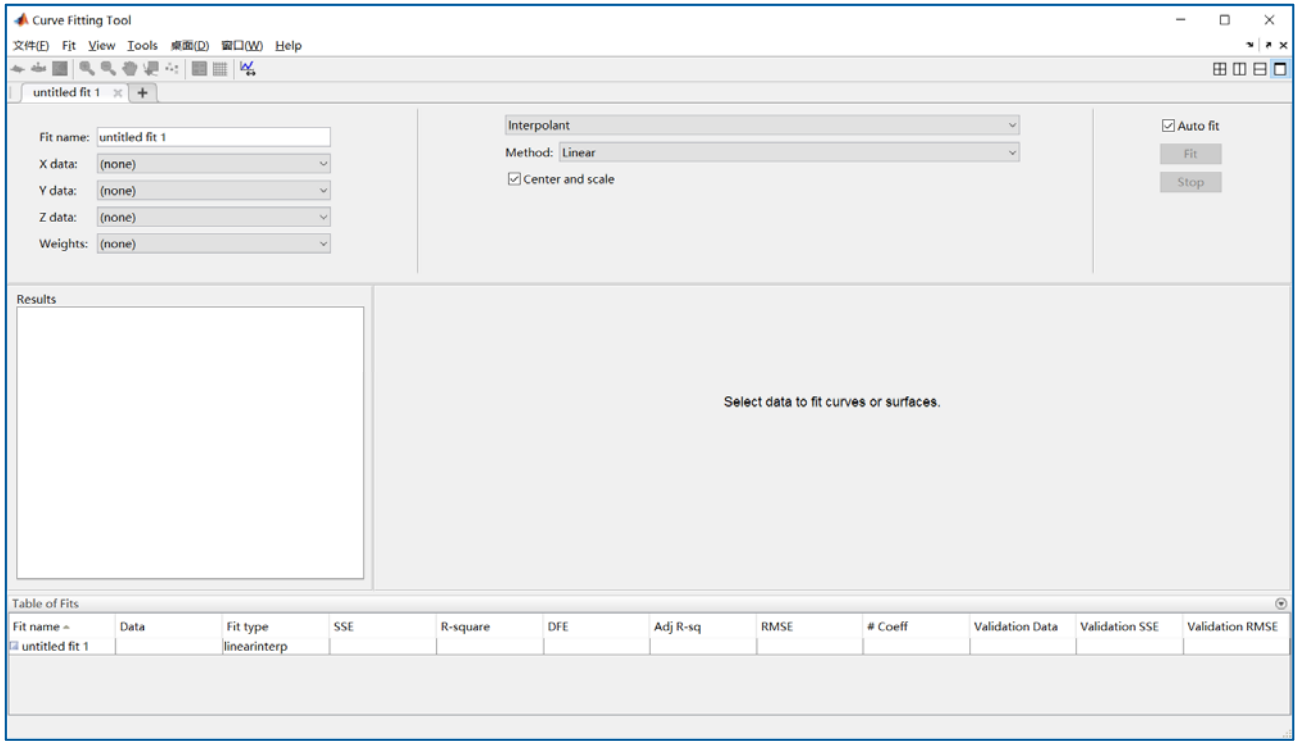
注:
mean()是求均值的函数。

```
SSR =

  151.1583


SSE =

    5.7281


SST =

  156.8863


ans =

   5.6843e-14


R_2 =

    0.9635
```

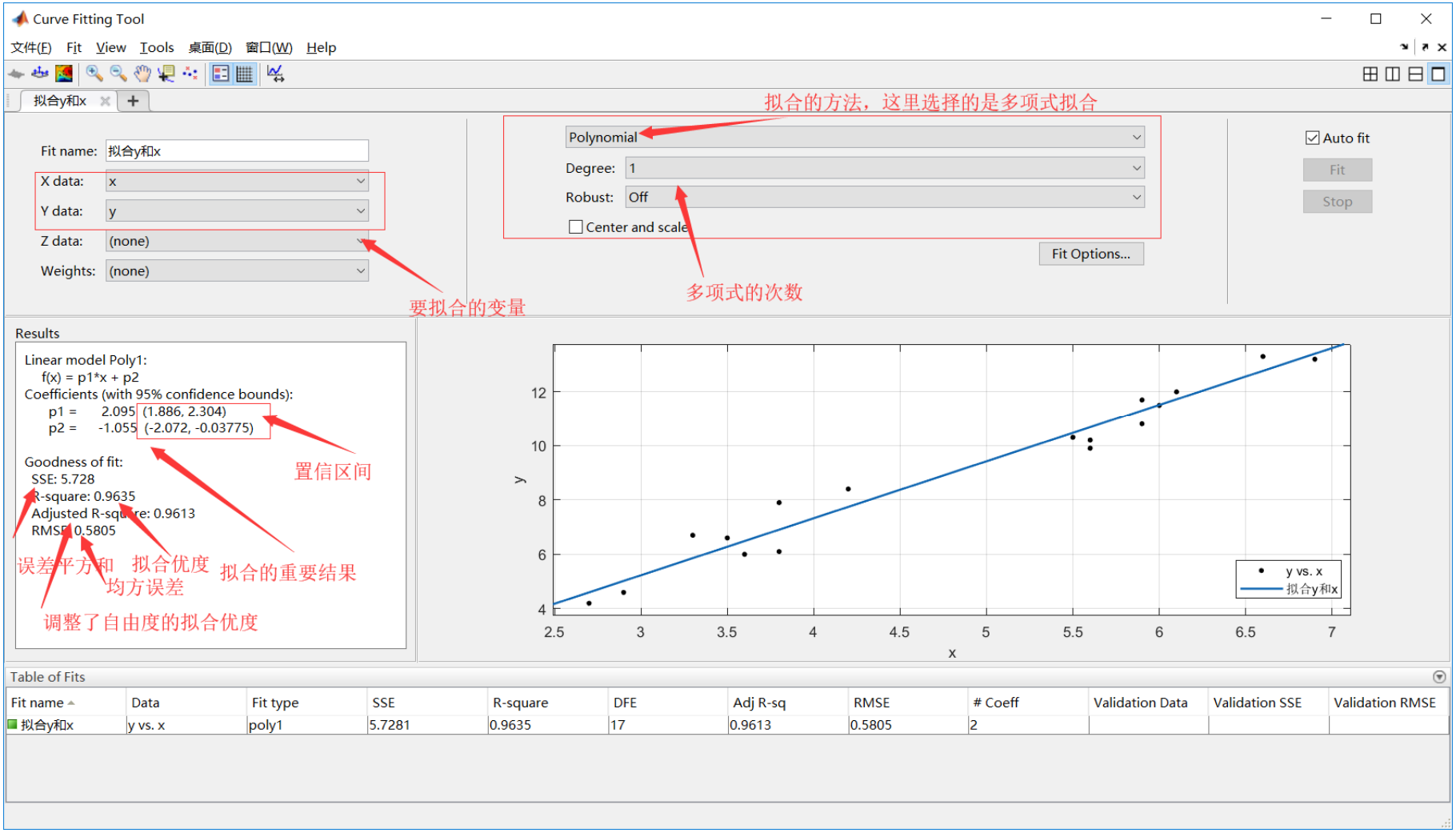
强大的曲线拟合工具箱



低版本的Matlab可以在命令窗口中直接输入"cftool"



拟合工具箱演示



利用拟合工具箱预测美国人口

population_predict.m

下表给出了近2个世纪的美国人口统计数据（单位：百万人），请使用最下面给定的拟合函数预测后30年的美国人口。

年	1790	1800	1810	1820	1830	1840	1850	1860
人口	3.9	5.3	7.2	9.6	12.9	17. 1	23.2	31.4
年	1870	1880	1890	1900	1910	1920	1930	1940
人口	38.6	50.2	62.9	76.0	92.0	106.5	123.2	131.7
年	1950	1960	1970	1980	1990	2000		
人口	150.7	179.3	204.0	226.5	251.4	281.4		

$$x(t)=\frac{x_m}{1+(\frac{x_m}{3.9}-1)e^{-r(t-1790)}}$$

x_m和r是两个拟合参数，t表示年份，$x(t)$表示第t年的人口

自己模拟数据进行演示

代码文件:code2.m

（1）randi：产生均匀分布的随机整数

```
%产生一个1至10之间的随机矩阵，大小为2x5;
s1 = randi(10,2,5);
%产生一个-5至5之间的随机矩阵，大小为1x10;
s2 = randi([-5,5],1,10);
```

（2） rand: 产生均匀分布的随机数

```
%产生一个0至1之间的随机矩阵，大小为1x5;
s3 = rand(1,5);
%产生一个a至b之间的随机矩阵，大小为1x5;
% a + (b-a) * rand(1,5); 如：a,b = 2,5
s4= 2 + (5-2) * rand(1,5);
```

（3）normrnd:产生正态分布的随机数

```
%产生一个均值为0，标准差为2的正态分布的随机矩阵，大小为3x4;
s5 = normrnd(0,2,3,4);
```

（4）roundn—任意位位置四舍五入

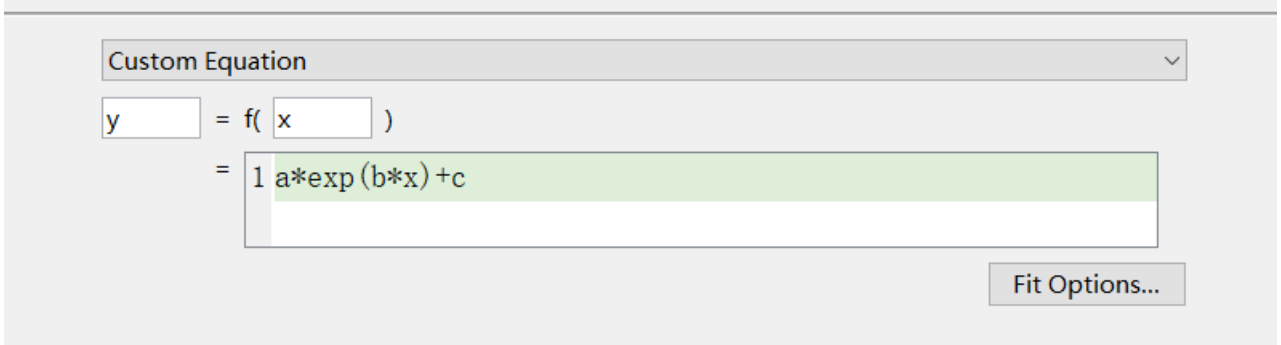
```
a = 3.1415
roundn(a,-2)     % ans    =  3.1400
roundn(a,2)       % ans    =  0
a =31415
roundn(a,2)    % ans   = 31400
```

自己模拟数据进行演示

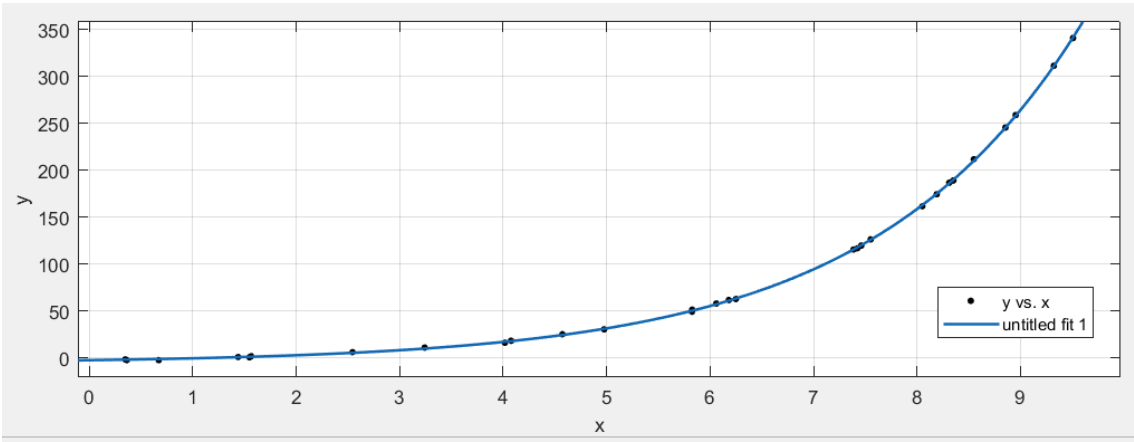
代码文件:code3.m

$$y_i = 3e^{0.5x_i} - 5 + \varepsilon_i \quad (i = 1, 2, \cdots, 30)$$

x_i是$[0, 10]$上的均匀分布，ε_i是标准正态分布的扰动项




```
General model:
    f(x) = a*exp(b*x)+c
Coefficients (with 95% confidence bounds):
       a =       3.089  (2.982, 3.196)
       b =      0.4966  (0.4929, 0.5003)
       c =      -5.195  (-5.859, -4.531)

Goodness of fit:
  SSE: 17.42
  R-square: 0.9999
  Adjusted R-square: 0.9999
  RMSE: 0.8032
```



优秀论文中的cftool运用

通过 MATLAB 分别利用指数函数、二次函数对两种机型风速与功率的实测数据进行拟合，拟合图像如图 4-6 和图 4-7（程序见附录 2.1 和 2.2）

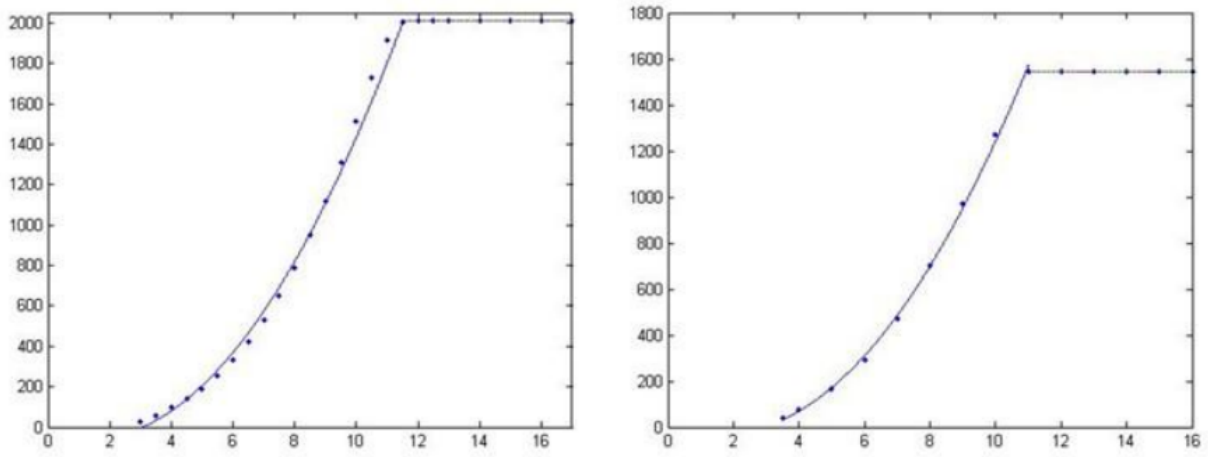


图4-6　两种机型风速与功率的二次函数拟合图

2.2　问题二中机型 II 的数据拟合

```
x=[3.5 4 5 6 7 8 9 10 11]
y=[40 74 164 293 471 702 973 1269 1544]
plot(x,y,'.')
hold on
x1=[3.5 4 5 6 7 8 9 10 11]
y1=[40 74 164 293 471 702 973 1269 1544]
cftool%此处调用工具箱分别对数据进行二次拟合与指数拟合
```

[2016年国赛高教杯奖D题]蚌埠士官学院-风电场运行状况分析及优化研究.pdf

优秀论文中的cftool运用

2.4 问题四的分析

建立鲢鱼、鳙鱼体长与体重之间关系的数学模型，考虑运用 matlab 中的 cftool 工具箱拟合曲线进行分析。再联系上一问“水华”爆发的时间和程度，根据鲢鱼和鳙鱼所食种类及其百分比和鱼类体重增长和消化食物的关系，建立放养数量的数学模型，得出需要在虾池中换养鱼的数量和时间，如果不能尽快消除“水华”，通过查阅文献寻找其他措施进行辅助。

我们由图像可大致观察出，随时间的增长，鲢鱼的体长和体重也在增加，且增加的速率越来越慢。我们可得出如下拟合曲线：

表 18　鲢鱼体长、体重增长方程

体长增长方程	拟合优度
$y=69.32e^{-(\frac{x-361.2}{267.9})^2}$	0.9704
$y=68.6\sin(0.004x+0.025)$	0.9967
体重增长方程	**拟合优度**
$y=2890e^{-(\frac{x-345.7}{306})^2}$	0.9512
$y=2860\sin(0.004x+0.186)$	0.982

注意：这里不要用R方，因为这些拟合函数不是线性函数，我们可以用SSE。

淡水养殖池塘华发生及自净化研究.pdf

cftool的‘骚’操作

更新13有三维图的绘制视频

针对问题一，对项目的任务定价规律进行**定性与定量**研究。利用 Matlab 的 cftool 工具箱绘制出任务的经纬度坐标与定价数据的**三维拟合图**，观察到任务分布密集的地区任务定价较低。对任务的位置数据进行**空间离散化处理**和 **K-Means 分析**，将任务分布的区域等划分为若干网格区域，定义影响任务定价的四个因子，即网格内任务数量、会员人数、会员平均完成能力、任务与中心点的距离。运用**灰色关联矩阵**定量分析四个**影响因子**与定价的相关度，分别为 0.9710, 0.9671, 0.9633, 0.9390。得出所定义的指标对定价相关性很高，能较好描述定价规律。最后通过比较未完成任务与已完成任务的相关度矩阵得出距离对任务的完成的影响是最显著的。

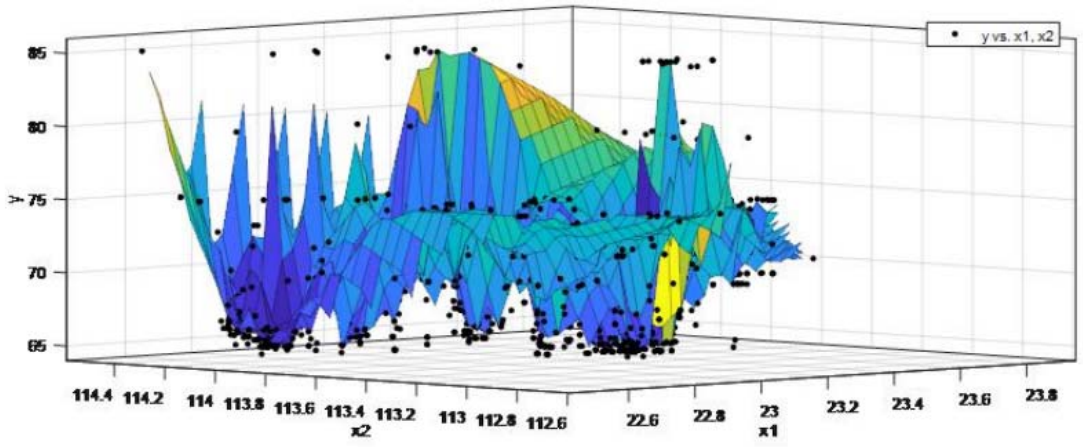


图 3 任务的经纬度及定价的三维曲线图

[2017年国赛高教杯奖B题]华中科技大学-基于聚类分析的双目标优化定价模型.pdf

课后作业

根据data2中的中国人口数据，确定你认为最合适的拟合函数，并说明原因。

	A	B
1	年份	人口(万)
2	2009	133126
3	2010	133770
4	2011	134413
5	2012	135069
6	2013	135738
7	2014	136427
8	2015	137122
9	2016	137866
10	2017	138639
11	2018	139538
12		